VISUAL OBJECT TRACKING IN DRONE IMAGES WITH DEEP REINFORCEMENT LEARNING

25th International Conference on Pattern Recognition

Derya GÖZEN Sedat OZER

Bilkent University

January 2021

Motivation

Goal:

- ✓ To develop a deep reinforcement learning (RL) based single object tracker for drone images

Main Contributions:

- ✓ introduction of a reinforcement learning based (RL) deep single object tracker for drone videos,
- ✓ introduction of a novel reward function for an improved performance for tracking objects in drone videos,
- ✓ introduction of new action types for drone data sets,
- ✓ testing our algorithm on two data sets: VisDrone2019 and OTB-100.

Deep RL-Based Trackers

- RL based deep trackers have been utilized on ground taken videos lately in the literature (see for example: ADNet*)
- We study the performance of RL based visual object trackers on drone videos and introduce four new RL based trackers. In those four models:
    - We study the effect of including **new actions**.
    - We study the effect of **architectural changes** in the model.
    - We study the effect of **reward function.**

* Yun, Sangdoo, et al. "Action-decision networks for visual tracking with deep reinforcement learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Deep RL-Based Trackers

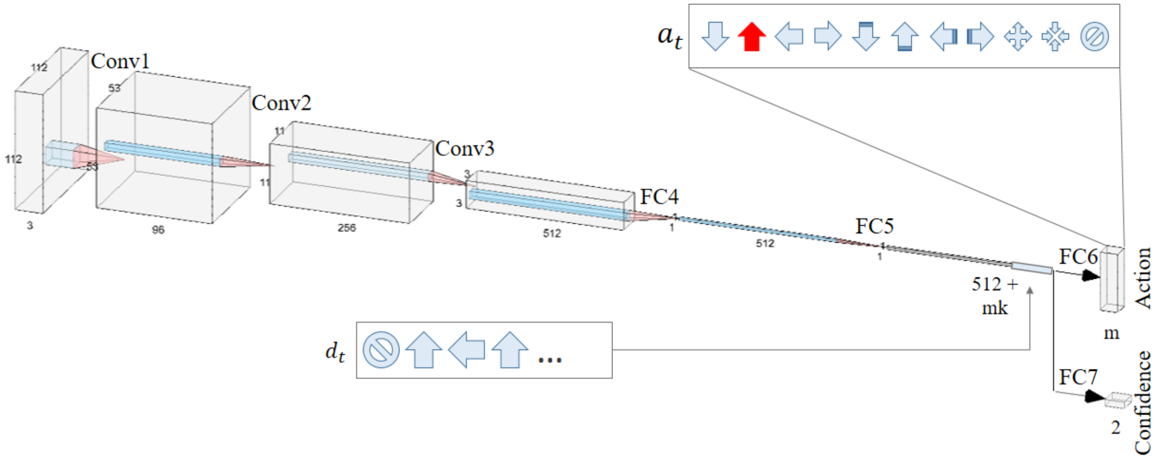


Fig. 3: Architecture of our baseline network, ADNet.

- The action-decision network, ADNet*, determines the future position of the object of interest in terms of a predicted action sequence, and each action in the sequence is predicted from the current state.
- A Markov Decision Process strategy for tracking
  - States $s \in S$
  - Actions a $\in A$
  - State transition function $s' = f(s, a)$
  - Reward function $r(s, a)$
- **Training**
  1. **Supervised learning – $\{\boldsymbol{w1}, \boldsymbol{w2}, \dots, \boldsymbol{w7}\}$**
  2. **Reinforcement learning – $\{\boldsymbol{w1}, \boldsymbol{w2}, \dots, \boldsymbol{w6}\}$**
  3. **Online adaptation in tracking – $\{\boldsymbol{w4}, \boldsymbol{w5}, \boldsymbol{w6}, \boldsymbol{w7}\}$**

* Yun, Sangdoo, et al. "Action-decision networks for visual tracking with deep reinforcement learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Our Action-Sequence-Based RL Trackers

Model A: Study the effect of using different actions

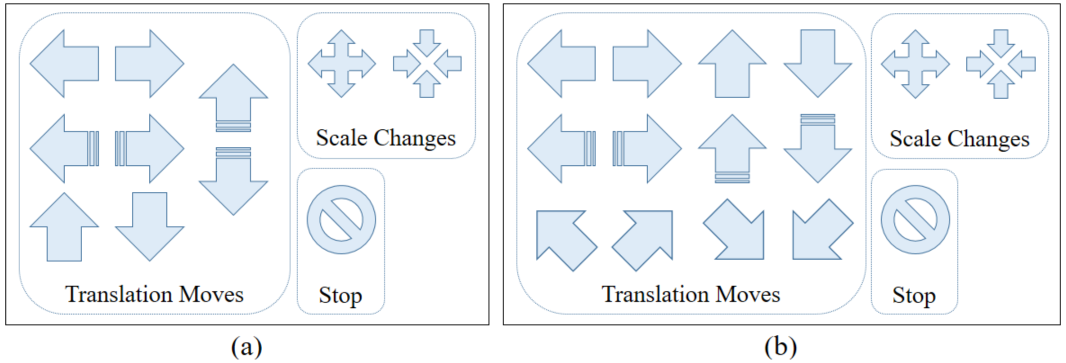


Fig. 4: The set of actions defined for (a) our base-network in ADNet, and for (b) *Model-A*.

- Action set utilization
  - 12 directional movements
  - 2 actions for scale changes
  - 1 terminal action (*stop*)

Our Action-Sequence-Based RL Trackers

Model B: Study the effect of changing backbone network

- Backbone network
  - ADNet uses VGG-M [4]
  - *Model-B* uses VGG-F [4]

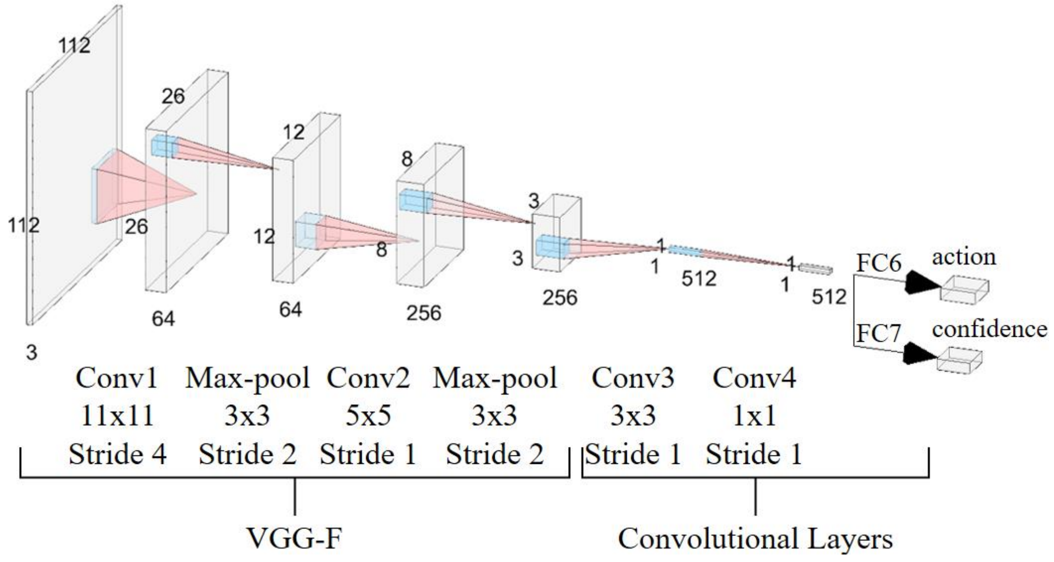


Fig. 5: Architecture of *Model-B*.

Our Action-Sequence-Based RL Trackers

Model C: Study the effect of reward function

Adjustments

Reward function	RL algorithm

Reward function

A hybrid reward function in the reinforcement learning stage, where the length of action set and the overlap ratio are both included during the rewarding process:

$$r(s_T) = \begin{cases} \left(10 - length(\{a_{t,l}\})\right) * IoU(b_T, G), & if\ IoU(b_T, G) > 0.70 \\ -1, otherwise & \end{cases}$$

RL algorithm

Data: Pre-trained network (W_{SL}), training sequences $\{F_l\}$ and ground truths $\{G_l\}$
Result: Trained network weights (W_{RL}) Initialize W_{RL} with W_{SL};
while W_{RL} *does not converge* do
    Randomly select $\{F_l\}_{l=1}^{L}$ and $\{G_l\}_{l=1}^{L}$
    Set initial $b_{1,1} \leftarrow G_1$
    Set initial $d_{1,1}$ as zero vector
    $T_1 \leftarrow 1$
    for l ← 2 to L do
        $\{a_{t,l}\}, \{b_{t,l}\}, \{d_{t,l}\}, T_l \leftarrow$ TRACKING($b_{T_{l-1},l-1}, d_{T_{l-1},l-1}, F_l$)
        Compute tracking scores $\{z_{t,l}\}$ with $\{b_{t,l}\}$ and $\{G_l\}$
        Calculate ΔW_{RL}
        Update W_{RL} using ΔW_{RL}
    end
end
Algorithm 1: Action-Sequence-Based Tracker (*Model-C*)

Our Action-Sequence-Based RL Trackers

Model D

Reward function

The reward function of our baseline network, ADNet:

$$r(s_T) = \begin{cases} 1, if\ IoU(b_T, G) > 0.70 \\ -1, otherwise \end{cases}$$

RL algorithm

Data: Pre-trained network (W_{SL}), training sequences $\{F_l\}$ and ground truths $\{G_l\}$
Result: Trained network weights (W_{RL}) Initialize W_{RL} with W_{SL};
while W_{RL} *does not converge* do
    Randomly select $\{F_l\}_{L l=1}$ and $\{G_l\}_{L l=1}$
    Set initial $b_{1,1} \leftarrow G_1$
    Set initial $d_{1,1}$ as zero vector
    $T_1 \leftarrow 1$
    for $l \leftarrow 2$ to L do
        $\{a_{t,l}\}, \{b_{t,l}\}, \{d_{t,l}\}, T_l \leftarrow$ TRACKING($b_{Tl-1,l-1}$, $d_{Tl-1,l-1}$, F_l)
        Compute tracking scores $\{z_{t,l}\}$ with $\{b_{t,l}\}$ and $\{G_l\}$
        Calculate ΔW_{RL}
        Update W_{RL} using ΔW_{RL}
    end
end
Algorithm 1: Action-Sequence-Based Tracker (*Model-C* and *Model-D*)

Used Data Sets

Data Sets

Ground-taken videos

- **Training data set**
  - ❖ VOT2013, VOT2014 and VOT2015 [2] (58 videos)
- **Test data set**
  - ❖ OTB-50 and OTB-100 [3] (100 videos)

Drone videos

- **Training data set**
  - ❖ VisDrone2019-SOT trainset part1 [4] (43 aerial videos)
- **Test data set**
  - ❖ VisDrone2019-SOT valset (11 aerial videos)

Fig. 1: Sample frames from VOT data sets.

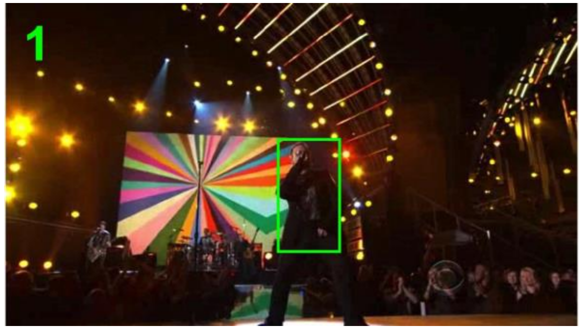
Fig. 2: Sample frames from VisDrone2019 data set.

Results & Analysis

Overall Performance

TABLE I: Comparison of our proposed methods to the baseline algorithm on OTB-100 and VisDrone2019 data sets.

Experiment Type	Model	OTB-100			VisDrone2019		
		Precision (20 pixels)	FPS	IoU	Precision (20 pixels)	FPS	IoU
Baseline model	ADNet	78.47%	4.89	0.603	89.15%	6.33	0.579
Action set	*Model-A*	79.45%	4.58	0.612	91.94%	6.08	0.557
Backbone network	*Model-B*	77.15%	8.11	0.574	89.67%	6.53	0.553
Reward function	*Model-C*	80.61%	6.25	0.589	93.02%	5.61	0.611
	Model-D	81.62%	7.02	0.616	91.74%	6.13	0.615



ADNet vs. ***Model-D*** on *Singer2*. Green, blue and red bounding boxes represent the ground truth, results of ADNet, and *Model-D*, respectively.

ADNet vs. ***Model-C*** on *uav0000092_00575_s*. Green, blue and red bounding boxes represent the ground truth, results of ADNet, and *Model-C*, respectively.

ADNet vs. ***Model-A*** on *uav0000317_02945_s*. Green, blue and red bounding boxes represent the ground truth, results of ADNet, and *Model-A*, respectively.

Results & Analysis

Challenging Aspects

ADNet vs. ***Model-D*** on *Skiing*. Green, blue and red bounding boxes represent the ground truth, results of ADNet, and *Model-D*, respectively.

ADNet vs. ***Model-D*** on *Panda*. Green, blue and red bounding boxes represent the ground truth, results of ADNet, and *Model-D*, respectively.

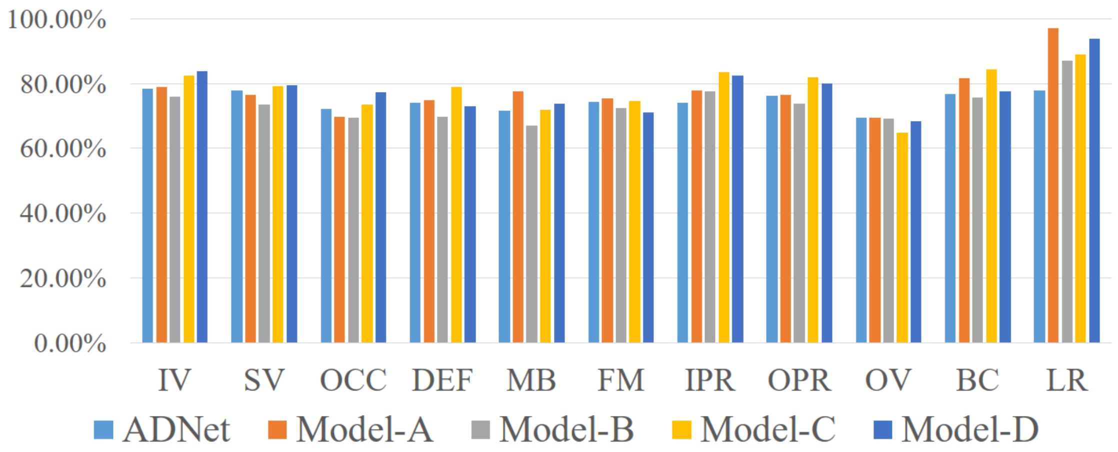


Fig. 4: Average precision results of ADNet, *Model-A*, *Model-B*, *Model-C*, and *Model-D* across the set of videos from OTB-100, grouped by challenging aspects.

THANK YOU!

ACKNOWLEDGMENT

This paper has been produced benefiting from the 2232 International Fellowship for Outstanding Researchers Program of TÜBİTAK (Project No: 118C356). However, the entire responsibility of the paper belongs to the owner of the paper. The financial support received from TÜBİTAK does not mean that the content of the publication is approved in a scientific sense by TÜBİTAK.

REFERENCES

[1] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2711–2720, 2017.

[2] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Ce- hovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–23, 2015.

[3] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013.

[4] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. Vision meets drones: Past, present and future. *arXiv preprint arXiv*:2001.06303, 2020.